CIS700: Reasoning for Natural Language Understanding

Project 1: Multi-hop Reading Comprehension through Question Decomposition and Rescoring by Min et al.

Krunal Shah

1 Observation

I reproduced the results of the authors and looking through the decomposed questions, I observed that a majority of the questions generated as a result of the decomposition phase did not make any sense. This meant that the subquestions obtained were not good enough to be justified as explainable evidence for the decision making of the model. The authors base their claim of the subquestions providing explainable evidence on the fact that their model generates subquestions that are almost as effective as human authored subquestions.

However, I understand the word "effective" to be key here. I observe that inspite of the subquestions not making sense as sentences, the words contained in the question did suggest what the question was trying to convey. Having previously worked with BERT model, I knew that it is robust to word eliminations and could work with missing words since it is able to capture the context from words with remarkable skill.

2 Motivation and Experiment Formulations

The above observations prompted me to try and understand how the model performance changes if the subquestions provided do not make any sense as questions or sentences. *The motivation behind this was to investigate if the subquestions generated by the model can indeed be viewed as explainable evidence or not.*

Towards this goal, I designed the following experimental settings:

1. *remove queries:* To make the subquestions lack the understanding of questions, I removed all words wh-words from the subquestions [1].

2. *reverse word order:* To make the subquestions lack the understanding of general sentences, I inverted the word order of the questions.

3 Results

The following results were obtained on the dev set of HotpotQA:

Setting	Overall F1	Bridge F1
Normal	70.045	71.858
Remove	69.670	71.353
Invert	68.567	70.065

[1] Specifically, the words 'what', 'which', 'who', 'when', 'where', 'whom', 'why', 'were' were removed

4 Conclusion

The results show that the performance drop of the model in the perturbation settings is minimal. This shows that the results of the model being as effective as human authored subquestions does not show that the subquestions are of high quality or even decent quality for that matter since we show that we can get almost as good results with subquestions which do not have any meaning as questions and sentences. This calls into question if the subquestions generated by the model can be counted as explainable evidence since if the subquestions do not necessarily represent sensible questions then the model still remains a black box and the subquestions do not shed any light on the model's reasoning behind its answers.